Journal of Statistical Planning and Inference 139 (2009) 2196 – 2207

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Bayesian likelihood robustness in linear models

Daniel Peña[a,*,1], Ruben Zamar[b,2], Guohua Yan[c,3]

[a]Departamento de Estadistica, Facultad de Ciencias Sociales, Universidad Carlos III de Madrid, Madrid 126, Getafe 28903, Spain
[b]Department of Statistics, University of British Columbia, 333-6356 Agricultural Road, Vancouver, British Columbia, Canada V6T 1Z2
[c]Department of Mathematics and Statistics, University of New Brunswick, Fredericton, New Brunswick, Canada E3B 5A3

ARTICLE INFO

Article history:
Received 6 October 2007
Received in revised form
6 August 2008
Accepted 15 October 2008
Available online 1 November 2008

Keywords:
Bayesian inference
Heteroscedasticity
Kullback–Leibler divergence
Robust regression

ABSTRACT

This paper deals with the problem of robustness of Bayesian regression with respect to the data. We first give a formal definition of Bayesian robustness to data contamination, prove that robustness according to the definition cannot be obtained by using heavy-tailed error distributions in linear regression models and propose a heteroscedastic approach to achieve the desired Bayesian robustness.

1. Introduction

Bayesian robustness comprises robustness to perturbation in the prior and in the data. As in Bayesian statistics the selection of the prior is an important and polemic topic, prior robustness has been the subject of much research. In prior robustness a set of prior distributions are considered and the range of a certain measure of interest when the prior varies over this class is studied. Less progress has been made regarding likelihood robustness. Box and Tiao (1968) in a seminal paper proposed robustifying the normal likelihood by considering that the data is generated by a mixture of two normals with different variances. They showed that the estimates obtained by this modeling procedure have some robustness properties. An alternative approach has been based on completely abandoning the normal likelihood and using heavy-tailed distributions for the noise. Some classic references on Bayesian robustness are Berger and Berliner (1986), Lavine (1991) and West (1984). More recent references on this topic are Abraham (2005), Bayarri and Berger (1998), Bayarri and Morales (2003), de Santis (2006) and Shyamalkumar (2000).

This paper has three main contributions. First, we introduce a formal definition of Bayesian robustness to data contamination. Second, we prove that this type of robustness cannot be obtained by using heavy-tailed error distributions in linear regression models. Third, we propose a heteroscedastic model to achieve the desired robustness in Bayesian linear regression models.

The paper is organized as follows. Section 2 presents a definition of likelihood robustness and shows that any Bayesian model that assume that the errors are independent and identically distributed cannot be robust according to this definition. An alternative approach is presented in Section 3 that leads to a modified likelihood that provides robust Bayesian inference according to our definition. Section 4 gives several examples to illustrate the performance of the proposed heteroscedastic approach. Section 5 concludes.

* Corresponding author. Tel.: +34 91 6249849.
E-mail addresses: dpena@est-econ.uc3m.es (D. Peña), ruben@stat.ubc.ca (R. Zamar), gyan@unb.ca (G. Yan).
[1] Supported by MEC Grant SEJ2007-64500.
[2] Supported by an NSERC discovery grant.
[3] Supported by a University of British Columbia Graduate Fellowship.

0378-3758/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2008.10.012

2. A definition of likelihood robustness

Suppose we have data $\mathbf{Z}$, that are assumed to be generated by some statistical model $f(\mathbf{Z}|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of unknown parameters. Let $\pi(\boldsymbol{\theta})$ be the prior and $p(\boldsymbol{\theta}|\mathbf{Z})$ be the joint posterior distribution for the vector of parameters $\boldsymbol{\theta}$ given the sample $\mathbf{Z}$. Let $p(\boldsymbol{\theta}|\mathbf{Z}_\alpha)$ be the posterior distribution for $\boldsymbol{\theta}$ given the sample $\mathbf{Z}_\alpha$ where a fraction α of points has been replaced by outliers of arbitrary size. Let $D(f_1, f_2)$ be a divergence function between the densities f_1 and f_2. We will say that the Bayesian inference is α-robust with respect to the divergence function D for the vector of parameters $\boldsymbol{\theta}$ if

$$\sup D(p(\boldsymbol{\theta}|\mathbf{Z}_\alpha), p(\boldsymbol{\theta}|\mathbf{Z})) < \infty,$$

where the supremum is taken over the set of all possible α-contaminated samples. Several possible alternatives for D can be found in Ullah (1996). The choice of the divergence D is important to obtain meaningful conclusions. We propose to use the Kullback–Leibler divergence, which is a common choice in Bayesian inference:

$$KL(\mathbf{Z}_\alpha, \mathbf{Z}) = \int \log\left(\frac{p(\boldsymbol{\theta}|\mathbf{Z}_\alpha)}{p(\boldsymbol{\theta}|\mathbf{Z})}\right) p(\boldsymbol{\theta}|\mathbf{Z}_\alpha)\, \mathrm{d}\boldsymbol{\theta},$$

which will be large when the contaminated posterior gives high density to parameter values which have very low density under the uncontaminated posterior.

A possible way to make the model robust in some cases may be to choose a prior distribution with compact support. Of course this is in general not compatible with true prior beliefs. A more general way is to set up the model so that the original prior and likelihood function are preserved but a mechanism is incorporated to down-weight discordant data points.

In this paper we will be mainly interested in the standard regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \tag{1}$$

where $\mathbf{y}$ and $\mathbf{u}$ are $n \times 1$, $\mathbf{X}$ is $n \times p$ and $\boldsymbol{\beta}$ is $p \times 1$. It is assumed that the observations are independent. Let $f(u)$ be the noise density, which is known up to a scale parameter σ, $f(u) = (1/\sigma)f_0(u/\sigma)$, $\pi(\boldsymbol{\beta}, \sigma^2)$ be the prior density and $p(\boldsymbol{\beta}, \sigma^2 |\ \mathbf{X}, \mathbf{y})$ be the posterior density. Then we have the following theorem.

Theorem 1. *Consider the regression model* (1) *satisfying the following assumptions*:

(i) *the noise density f is known up to a scale parameter* σ, $f(u) = (1/\sigma)f_0(u/\sigma)$, *where* f_0 *is continuous and* $uf_0(u) \to 0$ *as* $u \to \pm\infty$;
(ii) *the prior density* $\pi(\boldsymbol{\beta}, \sigma^2)$ *is continuous on its support and*
(iii) *the posterior density* $p(\boldsymbol{\beta}, \sigma^2|\mathbf{X}, \mathbf{y})$ *is proper and has finite number of modes*.

For any $\alpha > 0$, *a necessary condition for the Bayesian inference based on independent observations to be* α*-robust is that the function* $-\log(f_0(u))$ *is bounded above, that is, it verifies*

$$\lim_{|u|\to\infty} [-\log(f_0(u))] \leqslant k < \infty.$$

Remark 1. A corollary of this theorem is that if we use a likelihood function built as the product of independent observations with common density $f(u) = (1/\sigma)f_0(u/\sigma)$ the Bayesian inference is generally not α-robust with respect to the Kullback–Leibler divergence function. In fact, if $\rho(u) = -\log(f_0(u))$ is bounded above then

$$\rho(u) = -\log(f_0(u)) \leqslant k,$$

which implies

$$f_0(u) \geqslant \mathrm{e}^{-k},$$

and the integral of the density diverges. On the other hand, if $f_0(u)$ is a density, so that $\lim_{|u|\to\infty} f_0(u) \to 0$, then $\rho(u)$ cannot be bounded above. Thus any Bayesian robust procedure solely based on a heavy-tailed distribution cannot be α-robust with respect to the Kullback–Leibler divergence function.

Remark 2. Note that this theorem is quite inclusive. For the likelihood, it includes Gaussian distributions or heavy-tailed distributions such as t distribution, or finite mixture models as in Smith et al. (1996); for prior density, it includes proper or improper distributions as long as the posterior is proper.

Proof of Theorem 1. Without loss of generality, assume that the regression model does not include the intercept. Let us show that if $-\log(f_0(u))$ is not bounded above then the Bayesian inference on $(\boldsymbol{\beta}, \sigma^2)$ is not α-robust with respect to the Kullback–Leibler

divergence function. Given $\mathbf{Z}=(\mathbf{X},\mathbf{y})$, we denote the data without the last observation by $\mathbf{Z}_{-n}=(\mathbf{x}_1,\ldots,\mathbf{x}_{n-1},y_1,\ldots,y_{n-1})$. The posterior density for $(\boldsymbol{\beta},\sigma^2)$ is

$$p(\boldsymbol{\beta},\sigma^2|\mathbf{Z})=K(\mathbf{Z})p(\sigma^2|\mathbf{Z}_{-n})p(\boldsymbol{\beta}|\sigma^2,\mathbf{Z}_{-n})\left\{\frac{1}{\sigma}f_0\left(\frac{y_n-\mathbf{x}_n'\boldsymbol{\beta}}{\sigma}\right)\right\},$$

where $K(\mathbf{Z})$ is the normalizing constant which is finite by Assumption (iii) and does not depend on the model parameters. By Assumption (iii) again, the posterior density $p(\sigma^2|\mathbf{Z}_{-n})$ is proper; together with Assumptions (i) and (ii), the posterior conditional density $p(\boldsymbol{\beta}|\sigma^2,\mathbf{Z}_{-n})$ is continuous and bounded. We want to prove that by changing the single observation $(\mathbf{x}_n,y_n)$, the Kullback–Leibler divergence can be made arbitrarily large. We now replace the last observation $(\mathbf{x}_n,y_n)$ as follows. Let $\bar{\beta}_1$ be an arbitrary value. Let k be an arbitrary constant and take $\mathbf{Z}_\alpha=(\mathbf{Z}_{-n},\mathbf{x}_n^*,y_n^*)$ where $\mathbf{x}_n^*=(k,0,0,\ldots,0)$ and $y_n^*=k\bar{\beta}_1$. Then

$$p(\boldsymbol{\beta},\sigma^2|\mathbf{Z}_\alpha)=C_{k,\bar{\beta}_1}^{-1}p(\sigma^2|\mathbf{Z}_{-n})p(\boldsymbol{\beta}|\sigma^2,\mathbf{Z}_{-n})\left\{\frac{1}{\sigma}f_0\left(\frac{k(\bar{\beta}_1-\beta_1)}{\sigma}\right)\right\},$$

where

$$C_{k,\bar{\beta}_1}=\int p(\sigma^2|\mathbf{Z}_{-n})p(\boldsymbol{\beta}|\sigma^2,\mathbf{Z}_{-n})\left\{\frac{1}{\sigma}f_0\left(\frac{k(\bar{\beta}_1-\beta_1)}{\sigma}\right)\right\}\mathrm{d}\boldsymbol{\beta}\,\mathrm{d}\sigma^2.$$

We first show that $C_{k,\bar{\beta}_1}=\mathrm{O}(k^{-1})$. Let

$$z=\frac{k(\bar{\beta}_1-\beta_1)}{\sigma},$$

and

$$\boldsymbol{\beta}_{-1}=(\beta_2,\ldots,\beta_p).$$

Then $\beta_1=\bar{\beta}_1-\sigma z/k$ and

$$\begin{aligned}kC_{k,\bar{\beta}_1}&=\int p(\sigma^2|\mathbf{Z}_{-n})p\left(\bar{\beta}_1-\frac{\sigma z}{k},\boldsymbol{\beta}_{-1}|\sigma^2,\mathbf{Z}_{-n}\right)f_0(z)\,\mathrm{d}\boldsymbol{\beta}_{-1}\,\mathrm{d}z\,\mathrm{d}\sigma^2\\&=\int p(\sigma^2|\mathbf{Z}_{-n})p\left(\bar{\beta}_1-\frac{\sigma z}{k}|\sigma^2,\mathbf{Z}_{-n}\right)f_0(z)\,\mathrm{d}z\,\mathrm{d}\sigma^2.\end{aligned}$$

Since $p(\bar{\beta}_1-\sigma z/k|\sigma^2,\mathbf{Z}_{-n})$ is continuous and bounded and that

$$\int p(\sigma^2|\mathbf{Z}_{-n})f_0(z)\,\mathrm{d}z\,\mathrm{d}\sigma^2=1$$

by Lebesgue's dominated convergence theorem,

$$\begin{aligned}\lim_{k\to\infty}kC_{k,\bar{\beta}_1}&=\int\lim_{k\to\infty}p(\sigma^2|\mathbf{Z}_{-n})p\left(\bar{\beta}_1-\frac{\sigma z}{k}|\sigma^2,\mathbf{Z}_{-n}\right)f_0(z)\,\mathrm{d}z\,\mathrm{d}\sigma^2\\&=\int\lim_{k\to\infty}p(\sigma^2|\mathbf{Z}_{-n})p(\bar{\beta}_1|\sigma^2,\mathbf{Z}_{-n})f_0(z)\,\mathrm{d}z\,\mathrm{d}\sigma^2\\&=p(\bar{\beta}_1|\mathbf{Z}_{-n}).\end{aligned}$$

By Assumption (i), $uf_0(u)\to 0$ as $u\to\pm\infty$. Therefore,

$$\lim_{k\to\infty}p(\boldsymbol{\beta},\sigma^2|\mathbf{Z}_\alpha)=\begin{cases}\infty, & \beta_1=\bar{\beta}_1,\\ 0, & \beta_1\neq\bar{\beta}_1.\end{cases}$$

For any $\Delta>0$, let $D=\{(\boldsymbol{\beta},\sigma^2):|\bar{\beta}_1-\beta_1|\geqslant\Delta\}$. Following the same argument for $kC_{k,\bar{\beta}_1}$, we have

$$\begin{aligned}\lim_{k\to\infty}\mathrm{Prob}(D|\mathbf{Z}_\alpha)&=\lim_{k\to\infty}(kC_{k,\bar{\beta}_1})^{-1}\int_D p(\sigma^2|\mathbf{Z}_{-n})\lim_{k\to\infty}p\left(\bar{\beta}_1-\frac{\sigma z}{k},\boldsymbol{\beta}_{-1}|\sigma^2,\mathbf{Z}_{-n}\right)f_0(z)\,\mathrm{d}\boldsymbol{\beta}_{-1}\,\mathrm{d}z\,\mathrm{d}\sigma^2\\&=0.\end{aligned}$$

As a result, for a fixed $\bar{\beta}_1$, we are able to pick k such that

$$\int_{\{\beta_1:p(\beta_1|\mathbf{Z}_\alpha)\geqslant 1,\beta_1>\bar{\beta}_1-1\}}p(\beta_1|\mathbf{Z}_\alpha)\,\mathrm{d}\beta_1\geqslant\frac{1}{2}.\tag{2}$$

In fact, for $\Delta=\frac{1}{8}$, there exists k such that $\mathrm{Prob}(|\bar{\beta}_1-\beta_1|\geqslant\Delta|\mathbf{Z}_\alpha)<\frac{1}{4}$, whereas $\mathrm{Prob}(|\bar{\beta}_1-\beta_1|<\Delta,p(\beta_1|\mathbf{Z}_\alpha)\leqslant 1|\mathbf{Z}_\alpha)<\frac{1}{4}$.

Let us now show that the Kullback–Leibler divergence between $p(\beta_1|\mathbf{Z}_\alpha)$ and $p(\beta_1|\mathbf{Z})$ can be made arbitrarily large, i.e., for any $M>0$, there exists $\bar{\beta}_1$ and k such that

$$\int \log\left(\frac{p(\beta_1|\mathbf{Z}_\alpha)}{p(\beta_1|\mathbf{Z})}\right) p(\beta_1|\mathbf{Z}_\alpha)\,\mathrm{d}\beta_1 > M.$$

By Assumption (iii), the density function $p(\beta_1|\mathbf{Z})$ is proper with finite modes. Hence there exists $\bar{\beta}_1>0$ such that

$$p(\beta_1|\mathbf{Z}) < 1/\exp(2M+2)$$

for any β_1 with $\beta_1 > \bar{\beta}_1 - 1$. For this chosen $\bar{\beta}_1$, we pick k, which may depends on $\bar{\beta}_1$, such that Eq. (2) holds. Now we decompose the Kullback–Leibler divergence as

$$\begin{aligned}\int \log\left(\frac{p(\beta_1|\mathbf{Z}_\alpha)}{p(\beta_1|\mathbf{Z})}\right) p(\beta_1|\mathbf{Z}_\alpha)\,\mathrm{d}\beta_1 = &\int_{\{\beta_1:p(\beta_1|\mathbf{Z}_\alpha)<p(\beta_1|\mathbf{Z})\}} \log\left(\frac{p(\beta_1|\mathbf{Z}_\alpha)}{p(\beta_1|\mathbf{Z})}\right)\left(\frac{p(\beta_1|\mathbf{Z}_\alpha)}{p(\beta_1|\mathbf{Z})}\right) p(\beta_1|\mathbf{Z})\,\mathrm{d}\beta_1\\ &+\int_{\{(\beta_1):p(\beta_1|\mathbf{Z}_\alpha)\geqslant p(\beta_1|\mathbf{Z})\}} \log\left(\frac{p(\beta_1|\mathbf{Z}_\alpha)}{p(\beta_1|\mathbf{Z})}\right) p(\beta_1|\mathbf{Z}_\alpha)\,\mathrm{d}\beta_1.\end{aligned}$$

The first term is no less than $-\exp(-1)$ since $u\log(u)\geqslant -\exp(-1)$ for $0<u<1$. The second term is no less than

$$\begin{aligned}&\int_{\{\beta_1:p(\beta_1|\mathbf{Z}_\alpha)\geqslant p(\beta_1|\mathbf{Z}),p(\beta_1|\mathbf{Z}_\alpha)\geqslant 1,\beta_1>\bar{\beta}_1-1\}} \log\left(\frac{p(\beta_1|\mathbf{Z}_\alpha)}{p(\beta_1|\mathbf{Z})}\right) p(\beta_1|\mathbf{Z}_\alpha)\,\mathrm{d}\beta_1\\ &=\int_{\{\beta_1:p(\beta_1|\mathbf{Z}_\alpha)\geqslant 1,\beta_1>\bar{\beta}_1-1\}} \log\left(\frac{p(\beta_1|\mathbf{Z}_\alpha)}{p(\beta_1|\mathbf{Z})}\right) p(\beta_1|\mathbf{Z}_\alpha)\,\mathrm{d}\beta_1\\ &\geqslant\int_{\{\beta_1:p(\beta_1|\mathbf{Z}_\alpha)\geqslant 1,\beta_1>\bar{\beta}_1-1\}} \log\left(\frac{1}{1/\exp(2M+2)}\right) p(\beta_1|\mathbf{Z}_\alpha)\,\mathrm{d}\beta_1\\ &\geqslant M+1.\end{aligned}$$

Therefore, the Kullback–Leibler divergence between $p(\beta_1|\mathbf{Z}_\alpha)$ and $p(\beta_1|\mathbf{Z})$ can be made arbitrarily large by perturbing a single observation. This claim applies to the Kullback–Leibler divergence between $p(\boldsymbol{\beta},\sigma^2|\mathbf{Z}_\alpha)$ and $p(\boldsymbol{\beta},\sigma^2|\mathbf{Z})$ as well, by virtue of the following equality:

$$\begin{aligned}\int \log\left(\frac{p(\beta_1,\boldsymbol{\theta}|\mathbf{Z}_\alpha)}{p(\beta_1,\boldsymbol{\theta}|\mathbf{Z})}\right) p(\beta_1,\boldsymbol{\theta}|\mathbf{Z}_\alpha)\,\mathrm{d}\beta_1\,\mathrm{d}\boldsymbol{\theta} = &\int \log\left(\frac{p(\beta_1|\mathbf{Z}_\alpha)}{p(\beta_1|\mathbf{Z})}\right) p(\beta_1|\mathbf{Z}_\alpha)\,\mathrm{d}\beta_1\\ &+\int\left[\int \log\left(\frac{p(\boldsymbol{\theta}|\beta_1,\mathbf{Z}_\alpha)}{p(\boldsymbol{\theta}|\beta_1,\mathbf{Z})}\right) p(\boldsymbol{\theta}|\beta_1,\mathbf{Z}_\alpha)\,\mathrm{d}\boldsymbol{\theta}\right] p(\beta_1|\mathbf{Z}_\alpha)\,\mathrm{d}\beta_1,\end{aligned}$$

where $\boldsymbol{\theta}=(\boldsymbol{\beta}_{-1},\sigma^2)$. Therefore the inference is not α-robust with respect to the Kullback–Leibler divergence function. □

Remark 3. It is immediately from the proof of Theorem 1 that the mode of the posterior distribution using the contaminated data can be taken arbitrarily far away from that of the original data.

3. Robust regression-scale likelihood

From the Bayesian point of view it makes sense to express the uncertainty due to possible outliers in both the response and the carriers through the Bayesian modeling. This can be carried out as follows. We assume that the linear model applies in some range of the observed variables, that is, calling $\mathbf{z}=(\mathbf{y},\mathbf{x}_1,\ldots,\mathbf{x}_p)$ to the vector of variables and a norm $\|\mathbf{z}\|$, we assume that in the high density area of observations, defined by $\|\mathbf{z}\|\leqslant a$, we have

$$y_i=\mathbf{x}_i'\boldsymbol{\beta}+u_i, \tag{3}$$

and to allow for outliers in the response variable we assume a moderately heavy-tailed distribution for the noise, e.g. Laplace, $f(u)\propto(1/\sigma)\exp(-|u|/\sigma)$. Outside this region, for $|\mathbf{z}|\geqslant a$ we assume that (3) still holds, but subject to greater uncertainty represented by an increase in the noise variance, that is $f(u)\propto(w(\mathbf{z})/\sigma)\exp(-w(\mathbf{z})|u|/\sigma)$ where $w(\mathbf{z})\to 0$, as $|z|\to\infty$. A simple way to introduce the uncertainty is to assume that the variability increases with the distance from the center of the data. Then we can represent $w(\mathbf{z})$ with a Mahalanobis type of distance to the center of the data by using reliable estimates for the location and the scatter of the data. For simplicity we propose using the coordinate wise median, $\mathbf{m}$, the MAD (median of absolute deviations with respect to the median) and the quadrant correlation proposed by Huber (1981), which do not require great computational effort. The quadrant correlation between z_1 and z_2 is the ordinary correlation between the two variables $\mathrm{sign}(z_1-m_1)$ and $\mathrm{sign}(z_2-m_2)$. Let $\mathbf{R}$ be the quadrant correlation matrix for the $p+1$ variables and let $\mathbf{D}$ be the diagonal matrix with elements given by the

MAD of the $p+1$ variables. Define $\mathbf{C}=\mathbf{DRD}$. The uncertainty about the model in the sparse region will depend on

$$d_i=\sqrt{(\mathbf{z}_i-\mathbf{m})'\mathbf{C}^{-1}(\mathbf{z}_i-\mathbf{m})}, \tag{4}$$

$$a=\text{med}(d_i)+k\ \text{MAD}(d_i), \tag{5}$$

where $\text{med}(d_i)$ and $\text{MAD}(d_i)$ stand for median and MAD with respect to the median of the d_i values. The constant k is chosen as a function of α, that is, how much protection against outliers we want. Then, we propose

$$w_i(\mathbf{Z})=\begin{cases}1 & \text{for } d_i\leqslant a,\\ (1+d_i^2-a^2)^{-1/2} & \text{otherwise.}\end{cases} \tag{6}$$

In this model, the likelihood will be

$$l(\boldsymbol{\beta},\sigma^2|\mathbf{Z})\propto(\sigma)^{-n}\prod_{i=1}^{n}w_i(\mathbf{Z})\exp\left\{-\frac{1}{\sigma}\sum|y_i-x_i'\boldsymbol{\beta}|w_i(\mathbf{Z})\right\}.$$

We now show that this posterior leads to α-robust Bayesian inference with respect to the Kullback–Leibler divergence function.

Theorem 2. *Suppose the regression model* (3) *where the variance of the noise is constant for* $d_i\leqslant a$, *where* d_i *is given by* (4) *and is proportional to* $(1+d_i^2-a^2)$ *for a given by* (5), *otherwise. If the posterior* $p(\boldsymbol{\beta},\sigma^2|\mathbf{Z})$ *in this model is proper, the corresponding Bayesian inference is* α*-robust with respect to the Kullback–Leibler divergence function.*

Proof. Denote the original data, possibly after rearrangement of rows, by

$$\mathbf{Z}=\begin{bmatrix}\mathbf{Z}_1\\ \mathbf{Z}_2\end{bmatrix}=[\mathbf{X},\mathbf{y}],$$

where $\mathbf{Z}_1$, the bulk of the data, is contamination-free and $\mathbf{Z}_2$, $\alpha(<0.5)$ proportion of the data, is subject to arbitrary contamination. The contaminated data is denoted by

$$\mathbf{Z}_\alpha=\begin{bmatrix}\mathbf{Z}_1\\ \mathbf{Z}_{2,\alpha}\end{bmatrix}.$$

Let

$$\mathbf{Z}_\alpha^*=\text{diag}(w_1(\mathbf{Z}_\alpha),\ldots,w_n(\mathbf{Z}_\alpha))\mathbf{Z}_\alpha.$$

We now show that the rows in $\mathbf{Z}_\alpha^*$ are bounded: the Euclidean norm $\|\|$ of each row $\mathbf{z}_i^*$ of $\mathbf{Z}_\alpha^*$ is bounded by a constant $B(\mathbf{Z}_1)$ which is determined by the bulk of the data. To show this let $\mathbf{v}=(\mathbf{a},\mathbf{b})$, where $\mathbf{a}=(a_1,a_2,\ldots,a_{n_a})$, $\mathbf{b}=(b_1,b_2,\ldots,b_{n_b})$ and $n_a>n_b$. Since median and MAD both have a breakdown point of 0.5, we have

$$\min(\mathbf{a})\leqslant\text{med}(\mathbf{v})\leqslant\max(\mathbf{a}),$$

and

$$\min_{i\neq j}\{|a_i-a_j|\}\leqslant\text{MAD}(\mathbf{v})\leqslant\max_{i\neq j}\{|a_i-a_j|\}.$$

If we regard the robust location $\mathbf{m}$ as a function of $\mathbf{Z}_\alpha$, then $\|\mathbf{m}(\mathbf{Z}_\alpha)\|\leqslant B_m(\mathbf{Z}_1)$; if we regard the dispersion $\mathbf{C}$ as a function of $\mathbf{Z}_\alpha$, then the eigenvalues of $\mathbf{C}(\mathbf{Z}_\alpha)$ have a lower and upper bounds $B_l(\mathbf{Z}_1)$ and $B_u(\mathbf{Z}_1)$, respectively. Therefore,

$$\|\mathbf{z}_i-\mathbf{m}\|/\sqrt{B_u(\mathbf{Z}_1)}\leqslant d_i\leqslant\|\mathbf{z}_i-\mathbf{m}\|/\sqrt{B_l(\mathbf{Z}_1)}.$$

Hence the median and MAD of $\{d_i: i=1,\ldots,n\}$ are restricted by those d_i's of rows of $\mathbf{Z}_1$; $a=\text{med}(d_i)+k\text{MAD}(d_i)$ has an upper bound $B_a(\mathbf{Z}_1)$.

For a row $\mathbf{z}_i^*=w_i(\mathbf{Z}_\alpha)\mathbf{z}_i$ with $d_i>a$, we have

$$\begin{aligned}d_i^*&=\sqrt{(\mathbf{z}_i^*-\mathbf{m})'\mathbf{C}^{-1}(\mathbf{z}_i^*-\mathbf{m})}\\ &\leqslant\sqrt{(\mathbf{z}_i^*-w_i\mathbf{m})'\mathbf{C}^{-1}(\mathbf{z}_i^*-w_i\mathbf{m})}+(1-w_i)\sqrt{\mathbf{m}'\mathbf{C}^{-1}\mathbf{m}}\\ &\leqslant d_i(1+d_i^2-a^2)^{-1/2}+\|\mathbf{m}\|/\sqrt{B_l(\mathbf{Z}_1)}\\ &\leqslant\max\{1,a\}+B_m(\mathbf{Z}_1)/\sqrt{B_l(\mathbf{Z}_1)}.\end{aligned}$$

For a row $\mathbf{z}_i^*$ with $d_i \leqslant a$, we have $d_i^* \leqslant a$. Therefore, for any $\mathbf{z}_i^*$,

$$\|\mathbf{z}_i^*\} \leqslant \max\{1, B_a(\mathbf{Z}_1)\} + B_m(\mathbf{Z}_1)\,/\sqrt{B_l(\mathbf{Z}_1)} \equiv B(\mathbf{Z}_1).$$

In our treatment, the working likelihood is

$$\begin{aligned} l(\boldsymbol{\beta}, \sigma^2|\mathbf{Z}_\alpha) &\propto \prod_{i=1}^{n} \frac{w_i(\mathbf{Z}_\alpha)}{\sigma} f_0\left(\frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta})w_i(\mathbf{Z}_\alpha)}{\sigma}\right) \\ &\propto \prod_{i=1}^{n} \frac{1}{\sigma} f_0\left(\frac{y_i^* - (\mathbf{x}_i^*)'\boldsymbol{\beta}}{\sigma}\right) \\ &\propto l(\boldsymbol{\beta}, \sigma^2|\mathbf{Z}_\alpha^*), \end{aligned}$$

where $y_i^* = y_i w_i(\mathbf{Z}_\alpha)$ and $\mathbf{x}_i^* = \mathbf{x}_i w_i(\mathbf{Z}_\alpha)$. Therefore, our treatment is essentially equivalent to transforming $\mathbf{Z}_\alpha$ into $\mathbf{Z}_\alpha^*$ and then making usual Bayesian inference based on $\mathbf{Z}_\alpha^*$.

From Theorem 1, it suffices to show that $KL(\mathbf{Z}_\alpha^*, \mathbf{Z})$ is bounded. By assumption the posterior $p(\boldsymbol{\beta}, \sigma^2|\mathbf{Z}_\alpha^*)$ is proper for all $\mathbf{Z}_\alpha^*$. Therefore, $KL(\mathbf{Z}_\alpha^*, \mathbf{Z})$ is well defined. Let

$$g(\mathbf{Z}_\alpha^*) \equiv KL(\mathbf{Z}_\alpha^*, \mathbf{Z}) = \int \log\left(\frac{p(\boldsymbol{\beta}, \sigma^2|\mathbf{Z}_\alpha^*)}{p(\boldsymbol{\beta}, \sigma^2|\mathbf{Z})}\right) p(\boldsymbol{\beta}, \sigma^2|\mathbf{Z}_\alpha^*)\,\mathrm{d}\boldsymbol{\beta}\,\mathrm{d}\sigma^2.$$

Then $g(\mathbf{Z}) = 0$ and g is a continuous on a compact set. Hence the α-robustness with respect to the Kullback–Leibler divergence function is established. □

4. Examples

Theorem 1 implies that one discordant data point is sufficient to break down the Bayesian inference in a linear model with independent and identically distributed errors, even if the error term ε_i is modeled with a heavy-tailed distribution, such as the commonly used Student's t distribution. In this section, we compare the performance of Student's t regression with small degrees of freedom and our proposed heteroscedastic approach through several classical data sets. Throughout this section, noninformative prior is used, i.e., $\pi(\boldsymbol{\beta}, \sigma^2) \propto 1/\sigma^2$. And we are interested solely on regression coefficients, regarding the scale parameter σ as a nuisance.

4.1. Brain and body weights for 28 species

First we consider the brain and body weights data set (available in the package "MASS" of the R software with the names "Animals"), which is published in Rousseeuw (1987, p. 57) and is also analyzed by Salibian-Barrera (2000). It includes average brain and body weights for 28 species of land animals. A model considered is

$$\log(\text{Brain weight}) = \beta_1 + \beta_2 \log(\text{Body weight}) + u.$$

The left panel of Fig. 1 is the scatterplot of the log-transformed data. There are three points off the obvious linear pattern, which are species "dipliodocus", "triceratops" and "brachiosaurus", respectively. We shall call the data set without these three species "clean" data; the data of all 28 species "contaminated" data. It is well known that when the noise is normal the posterior distribution for the regression coefficients follows a multivariate t distribution and when the noise is Student's t this posterior has no explicit form but can be computed by Gibbs sampling (see Appendix A).

In the right panel of Fig. 1, we illustrate the evolution of the Kullback–Leibler divergences as the three outlying points in the original data are translated horizontally to the right. The horizontal axis is the increment of this translation. The vertical axis is the Kullback–Leibler divergence of the posterior distribution of the regression coefficients when a model is fitted to the modified data versus the posterior distribution of the regression coefficients when the same model is fitted to the clean data. The solid curve is for the normal regression model; the dotted curve is for independent Student's t regression model and the dashed curve is for our proposed method. It is clear that the Kullback–Leibler divergence increases quite fast for the normal regression model as the outliers are translated further away. For the independent Student's t regression model, we use 4 degrees of freedom as this is widely used in the literature. Smaller degrees of freedom behave similarly though are more numerically demanding. This model gains some robustness as we can see the increase of Kullback–Leibler divergence is modest when the data are contaminated not very heavily. But it increases fast as the contamination is heavy. The Bayesian analysis using Student's t errors is not robust according to the proposed definition in the sense that a few points can break down the inference. For the proposed down-weighting method, the Kullback–Leibler divergence is almost not affected as we are assigning smaller weights to these three points when they are translated further away.

The 95% credible intervals for the regression coefficients in several cases are listed in Table 1. We focus on the credible intervals for the slopes. The proposed approach effectively down-weights the three points and leads to credible intervals very

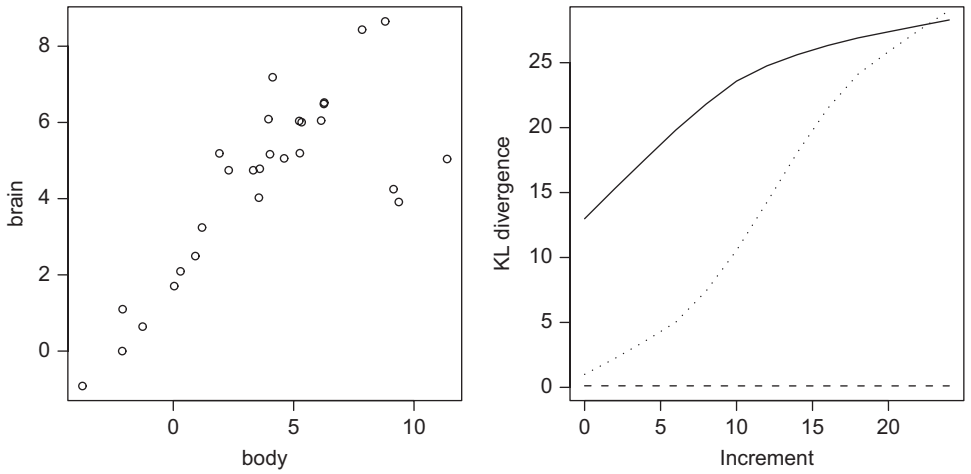

Fig. 1. Left panel: scatterplot of brain and body weights of 28 species. Right panel: Kullback–Leibler divergences curves of posterior distribution of regression coefficients versus that of clean data as the three outlying points are gradually translated to the right. Solid curve: normal regression; dashed curve: the proposed heteroscedastic approach and dotted curve: independent Student t regression with 4 degrees of freedom.

Table 1
The 95% credible intervals for the intercept and the slope in the brain and body weights data.

	Normal model (clean data)	Normal model	Proposed approach	t_4	t_1
Intercept	(1.72, 2.57)	(1.71, 3.41)	(1.77, 2.45)	(1.65, 2.83)	(1.65, 2.29)
Slope	(0.66, 0.85)	(0.33, 0.66)	(0.67, 0.83)	(0.52, 0.80)	(0.68, 0.82)

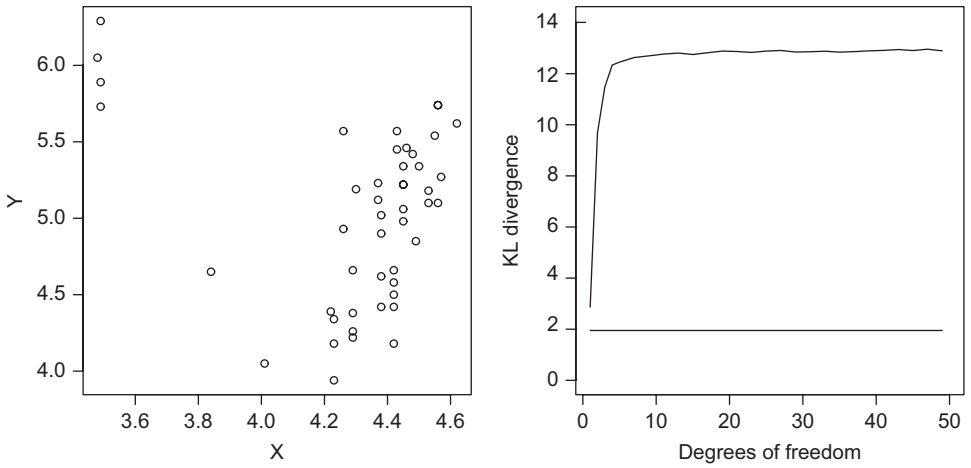

Fig. 2. Left panel: scatterplot of the star cluster CYG OB1. Right panel: the Kullback–Leibler divergence curve of the posterior distributions of the regression coefficients when independent Student t regression models are fitted to the original data and the clean data for various degrees of freedom. The horizontal line is the Kullback–Leibler divergence when the proposed approach is applied.

close to those from the clean data. In contrast, the normal regression model and the independent Student t regression models show the influence of these three points in various degrees and lead to wider or left-shifted credible intervals. The outliers have a catastrophic effect on the Bayesian approach with normal errors.

4.2. Stars cluster CYG OB1 data

This data set is also from Rousseeuw (1987, p. 27) and is available in the package "rrcov" of the R software. The data contain 47 stars in the direction of Cygnus. Here x is the logarithm of the effective temperature at the surface of the star, and y is the logarithm of its light intensity.

Table 2
The 95% credible intervals for the intercept and the slope in the stars cluster CYG OB1 data.

	Normal model (clean data)	Normal model	Proposed approach	t_4	t_1
Intercept	(−7.79, −0.37)	(4.36, 9.29)	(−12.59, −3.70)	(1.97, 9.38)	(−9.38, 8.48)
Slope	(1.20, 2.90)	(−1.00, 0.15)	(1.96, 3.98)	(−1.00, 0.69)	(−0.77, 3.26)

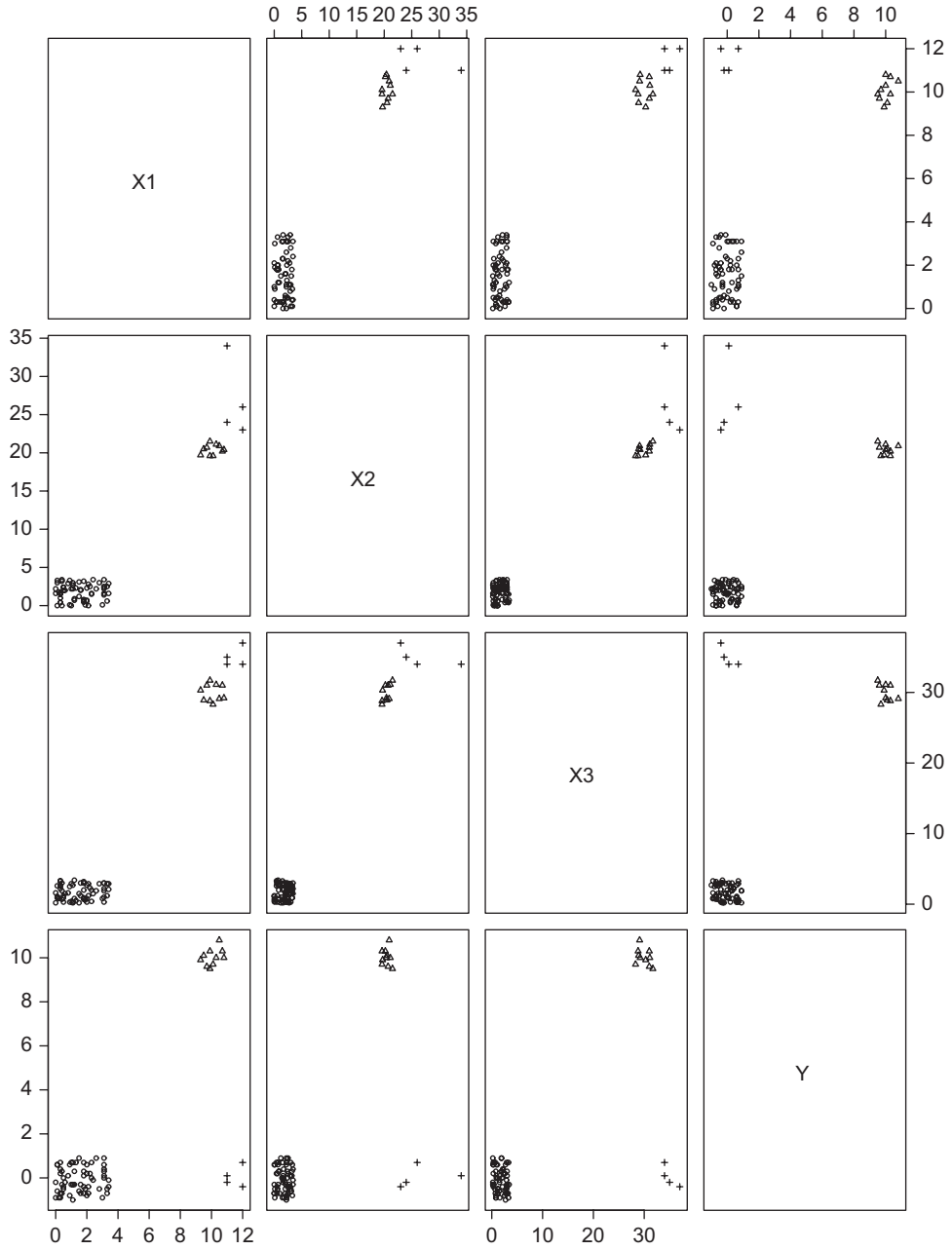

Fig. 3. Pairwise scatterplot of the Hawkins–Bradu–Kass data. Circles are points in the bulk, triangles are bad leverage points and plus signs are good leverage points.

In the scatterplot, left panel of Fig. 2, we can spot four outlying points in the upper corner which do not follow the general trend of the majority points. They are called giants (with indices 11, 20, 30 and 34); the remaining 43 points are said to lie on the main sequence. We call the data without the giants "clean" data.

Table 3
The 95% credible intervals for the intercept and the slopes in the Hawkins–Bradu–Kass data.

	Normal model (w/o obs 1–10)	Normal model	Proposed approach	t_4	t_1
β_0	$(-0.39, 0.03)$	$(-1.23, 0.44)$	$(-0.36, 0.33)$	$(-1.25, -0.60)$	$(-1.36, -0.32)$
β_1	$(-0.05, 0.21)$	$(-0.27, 0.77)$	$(-0.06, 0.19)$	$(-0.05, 0.34)$	$(-0.07, 0.33)$
β_2	$(-0.04, 0.12)$	$(-0.64, -0.02)$	$(-0.11, 0.14)$	$(-0.02, 0.33)$	$(0.03, 0.38)$
β_3	$(-0.12, 0.02)$	$(0.12, 0.64)$	$(-0.24, 0.01)$	$(0.07, 0.32)$	$(-0.08, 0.29)$

Table 4
Median weights of observations in the Hawkins–Bradu–Kass data from posterior sampling using the Bayesian approach with independent Student's t_1 errors.

Case no.	Med. wt.	Case no.	Med. wt.	Case no.	Med. wt.	Case no.	Med. wt.	Case no.	Med. wt.
1	1.00	16	0.49	31	1.15	46	1.06	61	0.55
2	0.79	17	0.78	32	1.22	47	0.73	62	0.14
3	1.03	18	1.27	33	0.42	48	1.18	63	1.23
4	0.23	19	1.22	34	1.12	49	0.50	64	0.26
5	0.72	20	1.14	35	1.21	50	0.40	65	0.39
6	1.03	21	0.40	36	0.24	51	0.26	66	0.70
7	0.36	22	1.01	37	0.45	52	0.99	67	0.60
8	0.38	23	0.43	38	0.15	53	0.24	68	0.15
9	0.51	24	0.30	39	1.11	54	0.35	69	0.40
10	0.98	25	0.64	40	1.10	55	1.02	70	0.50
11	0.00	26	0.24	41	0.85	56	1.21	71	1.31
12	0.00	27	0.21	42	0.80	57	0.20	72	1.30
13	0.00	28	0.57	43	0.45	58	0.76	73	0.50
14	0.00	29	0.36	44	0.60	59	0.75	74	0.48
15	0.35	30	1.10	45	1.12	60	0.38	75	0.46

We fit independent Student t regression models with various degrees of freedom to this data set. The right panel of Fig. 2 illustrates the evolution of Kullback–Leibler divergence of posterior distribution of regression coefficients on the original data versus that on clean data for various degrees of freedom. It is evident that the independent Student t regression model with lower degrees of freedom does gain some robustness. However, it is still not good enough since it gives rise to much wider credible intervals (see Table 2). Focusing on the slope, we might see that only the proposed approach leads to sensible inference close to the situation using the "clean" data and the corresponding credible interval indicate a significant positive slope.

4.3. Hawkins–Bradu–Kass data

This artificial data set was generated by Hawkins et al. (1984) and is available in the package "rrcov" of the R software as well. It consists of 75 observations in four dimensions (one response and three explanatory variables). The first 14 observations are outliers in four dimensions. In the regression sense, however, the first 10 observations are bad leverage points and the observations 11–14 are good leverage points. This data set provides a good example of the masking effect of multiple outliers. Bayesian unmasking is a hard problem (see for example Justel and Peña, 2001).

The pairwise scatterplot of the data is in Fig. 3. The 95% credible intervals of regression coefficients for several methods are shown in Table 3. We know that the intercept β_0 and slopes $\beta_1, \beta_2, \beta_3$ used to generate the data are all equal to zero. In the first column of Table 3, the 10 bad leverage points are not used when the normal regression model is fitted. All the credible intervals contain 0. These credible intervals serve as a gold standard for comparison. In the second column, when the normal regression is fitted to the whole data, credible intervals for β_2 and β_3 do not contain 0; when a Student's t regression model with 4 degrees of freedom is fitted to the whole data, credible intervals for β_0 and β_3 do not contain 0; when a Student's t regression model with 1 degrees of freedom is fitted, credible intervals for β_0 and β_2 do not contain 0. The proposed heteroscedastic approach down-weights all the 14 high leverage points and all resultant credible intervals contain 0, the true values of the parameters. Of course, the price to pay is that the 4 good leverage points are also down-weighted.

Furthermore, from Table 4 we see that the independent Student's t regression model with small degrees of freedom still cannot unmask the ten bad leverage points (observations 1–10) which are given weights similar to those of the majority of good data points. On the other hand, the four good leverage points (observations 11–14) are given very low weights and treated as regression outliers.

5. Conclusions

We have shown that Bayesian robustness according to our definition cannot be obtained by using a heavy-tailed distribution for independent errors. This result is consistent with previous results in the frequentist robustness literature (see for example, Yohai, 1987). To get around this negative result the frequentists build the estimator in two stages. First they compute a robust residual scale and second they use this robust scale to build a bounded loss function. However, in the frequentist approach

there is no formal model that can be associated with the inference process because the resulting bounded loss function does not correspond to any given likelihood. However, we have shown that it is possible to use a formal model for α-robust Bayesian inference with respect to the Kullback–Leibler divergence function, if we allow for heteroscedasticity directly determined by the data themselves.

Through several examples, we numerically compare the proposed approach with Bayesian approach with normal errors or heavy-tailed Student's t errors. In the brain and body weights data, we have shown that the Kullback–Leibler divergence is resistant to contamination for the proposed approach but not for the approaches with normal errors or heavy-tailed Student's t errors; and the other approaches lead to wider or shifted credible intervals. In the stars data, we have shown that only the proposed approach leads to sensible inference, that is, confirming the positive trend. In the Hawkins–Bradu–Kass data, we have shown that the proposed approach leads to credible intervals containing zero, which is the value used to generate the data, while its counterparts fail to unmask the bad leverage points. In conclusion, the proposed heteroscedastic approach works well as expected from the theory.

Appendix A. Computation of the Kullback–Leibler divergence

In the linear model (1), if we assume normal errors and use conjugate priors for $(\boldsymbol{\beta}, \sigma^2)$,

$$\pi(\boldsymbol{\beta}, \sigma^2) = \mathrm{N}(\boldsymbol{\beta}; b, \sigma^2 C^{-1})\mathrm{IG}(\sigma^2; \delta_1, \delta_2),$$

where IG indicates inverse Gamma distribution, then the posterior distribution of $(\boldsymbol{\beta}, \sigma^2)$ is still of the form

$$\pi(\boldsymbol{\beta}, \sigma^2|\mathbf{X}, \mathbf{y}) = \mathrm{N}(\boldsymbol{\beta}; b^*, \sigma^2 C^{*-1})\mathrm{IG}(\sigma^2; \delta_1^*, \delta_2^*),$$

where

$$b^* = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + C)^{-1}(\mathbf{X}^{\mathrm{T}}\mathbf{y} + Cb),$$

$$C^* = \mathbf{X}^{\mathrm{T}}\mathbf{X} + C,$$

$$\delta_1^* = \delta_1 + \frac{n-p}{2},$$

$$\delta_2^* = \delta_2 + \tfrac{1}{2}[\mathbf{y}^{\mathrm{T}}\mathbf{y} + b^{\mathrm{T}}Cb - (\mathbf{X}^{\mathrm{T}}\mathbf{y} + Cb)^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X} + C)^{-1}(\mathbf{X}^{\mathrm{T}}\mathbf{y} + Cb)].$$

We shall consider only the case of noninformative prior $\pi(\boldsymbol{\beta}, \sigma^2) \propto \sigma^{-2}$, which corresponds to $b = 0$, $C = 0$, $\delta_1 = 0$ and $\delta_2 = 0$. The posterior distribution of $(\boldsymbol{\beta}, \sigma^2)$ in this case is then

$$\pi(\boldsymbol{\beta}, \sigma^2|\mathbf{X}, \mathbf{y}) = \mathrm{N}(\boldsymbol{\beta}; \hat{\boldsymbol{\beta}}, \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1})\mathrm{IG}\left(\sigma^2; \frac{n-p}{2}, \frac{(n-p)\hat{\sigma}^2}{2}\right),$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$ and $\hat{\sigma}^2 = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})/(n-p)$. The marginal posterior distribution of $\boldsymbol{\beta}$ is multivariate t distribution

$$\boldsymbol{\beta}|\mathbf{X}, \mathbf{y} \sim T_{n-p}(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}).$$

The marginal likelihood is

$$m(\mathbf{y}|\mathbf{X}) = \frac{\Gamma\left(\frac{n-p}{2}\right)|\mathbf{X}^{\mathrm{T}}\mathbf{X}|^{-1/2}}{(\pi(n-p)\hat{\sigma}^2)^{n-p/2}}.$$

If

$$\pi(\boldsymbol{\beta}, \sigma^2|\mathbf{X}, \mathbf{y}) \sim \mathrm{N}(\boldsymbol{\beta}; b, \sigma^2 C^{-1})\mathrm{IG}(\sigma^2; \delta_1, \delta_2),$$

and

$$\pi(\boldsymbol{\beta}, \sigma^2|\mathbf{X}_\alpha, \mathbf{y}_\alpha) \sim \mathrm{N}(\boldsymbol{\beta}; b'^2(C'^{-1})\mathrm{IG}(\sigma^2; \delta_1', \delta_2'),$$

it is straightforward to show that the Kullback–Leibler divergence is

$$\begin{aligned}\mathrm{KL}(\pi(\boldsymbol{\beta}, \sigma^2|\mathbf{X}_\alpha, \mathbf{y}_\alpha), \pi(\boldsymbol{\beta}, \sigma^2|\mathbf{X}, \mathbf{y})) = {} & \log(|C'^{1/2}(\delta_2')^{\delta_1'}) - \log(|C|^{1/2}\delta_2^{\delta_1}) \\ & - \log(\Gamma(\delta_1')) + \log(\Gamma(\delta_1)) + (\delta_1' - \delta_1)(\Psi(\delta_1') - \log(\delta_2')) \\ & - \frac{1}{2}(p - \mathrm{tr}(C(C'^{-1}) - \left(b'^{\mathrm{T}}C(b' - b)\frac{\delta_1'}{\delta_2'}\right) - (\delta_2' - \delta_2)\frac{\delta_1'}{\delta_2'},\end{aligned}$$

where Ψ is the digamma function. In the case of noninformative priors, the Kullback-Leibler divergence is

$$\begin{aligned}\mathrm{KL}(\pi(\boldsymbol{\beta},\sigma^2|\mathbf{X}_\alpha,\mathbf{y}_\alpha),\pi(\boldsymbol{\beta},\sigma^2|\mathbf{X},\mathbf{y})) &= \frac{1}{2}\log(|\mathbf{X}_\alpha^{\mathrm{T}}\mathbf{X}_\alpha|/|\mathbf{X}^{\mathrm{T}}\mathbf{X}|)\\ &\quad+\frac{n-p}{2}\log(\hat{\sigma}_\alpha^2/\hat{\sigma}^2)+\frac{1}{2}(\mathrm{tr}((\mathbf{X}^{\mathrm{T}}\mathbf{X})^{\mathrm{T}}(\mathbf{X}_\alpha^{\mathrm{T}}\mathbf{X}_\alpha)^{-1})-p)\\ &\quad+\frac{(\hat{\boldsymbol{\beta}}_\alpha-\hat{\boldsymbol{\beta}})^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\hat{\boldsymbol{\beta}}_\alpha-\hat{\boldsymbol{\beta}})-(n-p)(\hat{\sigma}_\alpha^2-\hat{\sigma}^2)}{2\hat{\sigma}_\alpha^2}.\end{aligned} \tag{7}$$

If we are interested only in the regression coefficients $\boldsymbol{\beta}$, the Kullback–Leibler divergence can be approximated from posterior sampling. Let $\boldsymbol{\beta}^{(1)},\ldots,\boldsymbol{\beta}^{(M)}$ be a sample from $T_{n-p}(\hat{\boldsymbol{\beta}},\hat{\sigma}^2(\mathbf{X}_\alpha^{\mathrm{T}}\mathbf{X}_\alpha)^{-1})$. Then

$$\mathrm{KL}(\pi(\boldsymbol{\beta}|\mathbf{X}_\alpha,\mathbf{y}_\alpha),\pi(\boldsymbol{\beta}|\mathbf{X},\mathbf{y}))\approx\frac{1}{M}\sum_{m=1}^{M}\log\left\{\frac{T_{n-p}(\boldsymbol{\beta}^{(m)};\hat{\boldsymbol{\beta}}_\alpha,\hat{\sigma}_\alpha^2(\mathbf{X}_\alpha^{\mathrm{T}}\mathbf{X}_\alpha)^{-1})}{\mathrm{T}_{n-p}(\boldsymbol{\beta}^{(m)};\hat{\boldsymbol{\beta}},\hat{\sigma}^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1})}\right\}. \tag{8}$$

When we model the errors with independent Student's t distribution with fixed small v degrees of freedom, still assuming noninformative prior $\pi(\boldsymbol{\beta},\sigma^2)\propto 1/\sigma^2$, no analytical form of Kullback–Leibler distance is available. To approximate the Kullback–Leibler divergence, we need to compute marginal likelihoods $m(\mathbf{y}|\mathbf{X})$ and $m(\mathbf{y}_\alpha|\mathbf{X}_\alpha)$. Let $(\boldsymbol{\beta}^{(m)},(\sigma^2)^{(m)})$, $m=1,\ldots,M$ be an MCMC sample from $\pi(\boldsymbol{\beta},\sigma^2|\mathbf{X}_\alpha,\mathbf{y}_\alpha)$ after a burn-in period, then

$$\mathrm{KL}(\pi(\boldsymbol{\beta},\sigma^2|\mathbf{X}_\alpha,\mathbf{y}_\alpha),\pi(\boldsymbol{\beta},\sigma^2|\mathbf{X},\mathbf{y}))\approx\frac{1}{M}\sum_{m=1}^{M}\log\left\{\frac{m(\mathbf{y}_\alpha|\mathbf{X}_\alpha)^{-1}p(\mathbf{y}_\alpha|\mathbf{X}_\alpha,\boldsymbol{\beta}^{(m)},(\sigma^2)^{(m)})}{m(\mathbf{y}|\mathbf{X})^{-1}p(\mathbf{y}|\mathbf{X},\boldsymbol{\beta}^{(m)},(\sigma^2)^{(m)})}\right\}. \tag{9}$$

Newton and Raftery (1994) use the harmonic mean of likelihood values to estimate marginal likelihood, which is in our case,

$$m(\mathbf{y}_\alpha|\mathbf{X}_\alpha)\approx\left\{\frac{1}{M}\sum_{m=1}^{M}\frac{1}{p(\mathbf{y}|\mathbf{X},\boldsymbol{\beta}^{(m)},(\sigma^2)^{(m)})\pi(\boldsymbol{\beta}^{(m)},(\sigma^2)^{(m)})}\right\}^{-1}.$$

This is a direct use of the identity

$$m(y)=\frac{p(y|\theta)\pi(\theta)}{\pi(\theta|y)},$$

which Chib (1995) termed as the *basic marginal likelihood identity* (*BMI*). The harmonic mean approach is not stable as the inverse of likelihood does not have finite variance. Chib (1995) and Chib and Jeliazkov (2001, 2005) proposed alternatives catered to various posterior sampling strategies. Notice that the BMI is invariant to θ, Chib estimates $\pi(\theta^*|y)$ for a point θ^* with high density such as posterior mean or posterior median.

In independent Student's t regression model, the hierarchical modeling strategy makes Gibbs sampling possible (Geweke, 1993):

$$y_i=\mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta}+u_i,\quad u_i\sim\mathrm{N}(0,\sigma^2/\omega_i),\quad \omega_i\sim\Gamma\left(\frac{v}{2},\frac{v}{2}\right).$$

Let $\boldsymbol{\omega}=(\omega_1,\ldots,\omega_n)^{\mathrm{T}}$. Suppose $(\boldsymbol{\beta}^{(m)},(\sigma^2)^{(m)},\boldsymbol{\omega}^{(m)})$, $m=1,\ldots,M$ be a Gibbs sample from $\pi(\boldsymbol{\beta},\sigma^2,\boldsymbol{\omega}|\mathbf{X},\mathbf{y})$ after a burn-in period, then

$$\pi(\boldsymbol{\beta}^*|\mathbf{X},\mathbf{y})\approx\frac{1}{M}\sum_{m=1}^{M}\pi(\boldsymbol{\beta}^*|(\sigma^2)^{(m)},\boldsymbol{\omega}^{(m)},\mathbf{X},\mathbf{y}).$$

Let $(\sigma^2)_{\boldsymbol{\beta}^*}^{(m)},\omega_{\boldsymbol{\beta}^*}^{(m)})$, $m=1,\ldots,M$ be a Gibbs sample from the reduced posterior $\pi(\sigma^2,\boldsymbol{\omega}|\boldsymbol{\beta}^*,\mathbf{X},\mathbf{y})$, then

$$\pi((\sigma^2)^*|\boldsymbol{\beta}^*,\mathbf{X},\mathbf{y})\approx\frac{1}{M}\sum_{m=1}^{M}p((\sigma^2)^*|\boldsymbol{\beta}^*,\omega_{\boldsymbol{\beta}^*}^{(m)},\mathbf{X},\mathbf{y}).$$

Then the marginal likelihood

$$m(\mathbf{y}|\mathbf{X})\approx\frac{p(\mathbf{y}|\mathbf{X},\boldsymbol{\beta}^*,(\sigma^2)^*)\pi(\boldsymbol{\beta}^*,(\sigma^2)^*)}{\pi(\boldsymbol{\beta}^*|\mathbf{X},\mathbf{y})\pi((\sigma^2)^*|\boldsymbol{\beta}^*,\mathbf{X},\mathbf{y})}.$$

If we are interested in the regression coefficients $\boldsymbol{\beta}$ only, then

$$\mathrm{KL}(\pi(\boldsymbol{\beta}|\mathbf{X}_\alpha,\mathbf{y}_\alpha,\pi(\boldsymbol{\beta}|\mathbf{X},\mathbf{y}))) \approx \frac{1}{M}\sum_{m=1}^{M}\log\left\{\frac{\pi(\boldsymbol{\beta}^{(m)}|\mathbf{X}_\alpha,\mathbf{y}_\alpha)}{\pi(\boldsymbol{\beta}^{(m)}|\mathbf{X},\mathbf{y})}\right\}, \tag{10}$$

where $\boldsymbol{\beta}^{(m)}, m = 1,\ldots,M$ are a Gibbs sample from $\pi(\boldsymbol{\beta},\sigma^2,\boldsymbol{\omega}|\mathbf{X}_\alpha,\mathbf{y}_\alpha)$.

References

Abraham, C., 2005. Asymptotics in Bayesian decision theory with applications to global robustness. J. Multivariate Anal. 95 (1), 50–65.
Bayarri, M., Berger, J., 1998. Robust Bayesian analysis of selection models. Ann. Statist. 26 (2), 645–659.
Bayarri, M., Morales, J., 2003. Bayesian measures of surprise for outlier detection. J. Statist. Plann. Inference 111, 3–22.
Berger, J., Berliner, L., 1986. Robust Bayes and empirical Bayes analysis with ε-contaminated priors. Ann. Statist. 14 (2), 461–486.
Box, G., Tiao, G., 1968. A Bayesian approach to some outlier problems. Biometrika 55 (1), 119–129.
Chib, S., 1995. Marginal likelihood from the Gibbs output. J. Amer. Statist. Assoc. 90 (432), 1313–1321.
Chib, S., Jeliazkov, I., 2001. Marginal likelihood from the Metropolis–Hastings output. J. Amer. Statist. Assoc. 96 (453), 270–281.
Chib, S., Jeliazkov, I., 2005. Accept–reject Metropolis–Hastings sampling and marginal likelihood estimation. Statist. Neerlandica 59 (1), 30–44.
de Santis, F., 2006. Sample size determination for robust Bayesian analysis. J. Amer. Statist. Assoc. 101 (473), 278–291.
Geweke, J., 1993. Bayesian treatment of the independent Student-t linear model. J. Appl. Econom. 8, 19–40.
Hawkins, D., Bradu, D., Kass, G., 1984. Location of several outliers in multiple-regression data using elemental sets. Technometrics 26 (3), 197–208.
Huber, P., 1981. Robust Statistics. Wiley Series in Probability and Mathematical Statistics. Wiley, New York.
Justel, A., Peña, D., 2001. Bayesian unmasking in linear models. Comput. Statist. Data Anal. 36 (1), 69–84.
Lavine, M., 1991. An approach to robust Bayesian analysis for multidimensional parameter spaces. J. Amer. Statist. Assoc. 86 (414), 400–403.
Newton, M., Raftery, A., 1994. Approximate Bayesian inference with the weighted likelihood bootstrap. J. Roy. Statist. Soc. Ser. B (Methodological) 56 (1), 3–48.
Rousseeuw, P., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65.
Salibian-Barrera, M., 2000. Contributions to the theory of robust inference. Ph.D. Thesis, The University of British Columbia.
Shyamalkumar, N., 2000. Likelihood robustness. In: Rìos Insua, D., Ruggeri, F. (Eds.), Robust Bayesian Analysis. Springer, New York.
Smith, M.N., Sheather, S., Kohn, R., 1996. Finite sample performance of robust Bayesian regression. SSRN eLibrary.
Ullah, A., 1996. Entropy, divergence and distance measures with econometric applications. J. Statist. Plann. Inference 49, 137–162.
West, M., 1984. Outlier models and prior distributions in Bayesian linear regression. J. Roy. Statist. Soc. Ser. B 46 (3), 431–439.
Yohai, V., 1987. High breakdown-point and high efficiency robust estimates for regression. Ann. Statist. 15 (2), 642–656.